Robot Reinforcement Learning using EEG-based reward signals

I. Iturrate, L. Montesano, J. Minguez

Abstract— **Reinforcement learning algorithms have been successfully applied in robotics to learn how to solve tasks based on reward signals obtained during task execution. These reward signals are usually modeled by the programmer or provided by supervision. However, there are situations in which this reward is hard to encode, and so would require a supervised approach of reinforcement learning, where a user directly types the reward on each trial. This paper proposes to use brain activity recorded by an EEG-based BCI system as reward signals. The idea is to obtain the reward from the activity generated while observing the robot solving the task. This process does not require an explicit model of the reward signal. Moreover, it is possible to capture subjective aspects which are specific to each user. To achieve this, we designed a new protocol to use brain activity related to the correct or wrong execution of the task. We showed that it is possible to detect and classify different levels of error in single trials. We also showed that it is possible to apply reinforcement learning algorithms to learn new similar tasks using the rewards obtained from brain activity.**

I. Introduction

Robot learning covers a field of robotics where robots learn new abilities or improve their performance based on data related to the task. Examples of these techniques include imitation learning [1], where the robot learns from a demonstration, or learning through experience. In the latter case, the robot learns by acting and using the information provided by its actions to improve its knowledge about the environment. In this context, reinforcement learning methods (RL) [2] have been successfully applied to learn motor behaviors and motion primitives from reward signals obtained while acting. Furthermore, recent developments have made possible to apply reinforcement learning in real robot problems, where one has to cope with continuous states and spaces and many degrees of freedom e.g., [3].

The key ingredient of any RL method is to compute a policy that maximizes a reward signal (or minimize a cost). While acting, the robot receives samples of this reward and uses them to improve its own policy in the future. In practice, this reward signal is defined by the robot programmer for each specific task. To compute the particular reward, one has to develop some ad-hoc engineered system (e.g. a tracking system) or, alternatively, supervise the task and manually provide a reward signal.

This paper addresses a novel approach to compute the rewards for the learning task directly from brain activity recorded using a non-invasive Brain-Computer Interface. The long term vision is to develop robotic systems such as prostheses that operate close to the human and can adapt themselves to new tasks. By extracting the reward directly from brain activity, this adaptation process has several advantages. It occurs in a transparent manner even in situations where it would be difficult to model the reward. Furthermore, this process captures task subjective aspects that depend on each user, which is a firm step towards the individualized operation.

Brain-Computer Interfaces (BCI) are systems that record and process the brain activity to perform useful actions in the logic or physical world. The recording technique used in this paper is the electroencephalogram (EEG), which is a non-invasive method (it registers electrical activity on the surface of the scalp). Despite its low spatial resolution, the EEG is portable and has a very high temporal resolution. Therefore, it turns to be interesting for real-time applications and in particular for the field of robotics. EEG-based BCI systems have been used to move an arm prosthesis [4], drive a robotic wheelchair [5], [6], or teleoperate a robot via internet [7].

Broadly speaking, there are two types of brain activity: spontaneous brain rhythms and event-related potentials (ERPs). The difference is that the ERPs are evoked by stimuli or events (as opposed to the spontaneous EEG rhythms) and display stable time relationships to a determined reference event [8]. Due to these properties, the ERPs turn to be the natural choice for a robot learning setting in which the robot executes actions observed by the human that will elicit the ERP activity.

The important question is whether this ERP activity, originated while observing the robot executing a task, actually encodes useful information to evaluate the task, i.e to compute a reward. In this direction, in cognitive neuroscience and neurophysiopsychology it is well known the usage of the ERP to study the underlying mechanisms of the human error processing (recently aglutinated as ErrPs, see [9] and references therein). Different ErrPs have been described, for instance, when a subject performs a choice reaction task under time pressure and realizes that he has committed an error [10] (response ErrPs); when the subject is given feedback indicating that he has committed an error [11] (feedback or reinforcement ErrPs); when the subject perceives an error committed by another person (observation ErrPs) [12]; or when the subject delivers an order and the machine executes another one [9] (interaction ErrP). In addition to this, several works have shown that it is possible to use signal processing and machine learning techniques to perform automatic single trial classification of these ErrPs [9], [13].

Iñaki Iturrate, Luis Montesano and Javier Mínguez are with the Instituto de Investigación en ingeniería de Aragón (I3A) and Dpto. de Informática e Ingeniería de Sistemas (DIIS), Universidad de Zaragoza, Spain. E-mail: iturrate@unizar.es, montesano@unizar.es and jminguez@unizar.es.
This work has been partially supported by projects HYPER-CSD2009-00067, DPI2009-14732-C02-01 funded by the Spanish Government and the Portuguese FCT project PTDC/EEA-ACR/70174/2006.

All the previous research forms the basis for the automatic computation of rewards for learning tasks based on EEG brain activity. To our knowledge, there is only a recent paper addressing a similar problem where brain activity is used during a learning process [14]. This rather preliminary study directly modified the probabilities of a policy in a task with two actions and two states i.e., a two parameter policy. Whenever an error was detected, the probability of the corresponding action was decreased using the entropy of the policy. The contribution of this paper is to push forward the understanding of how ERPs can be used in RL algorithms for robot learning. A new protocol to elicit event-related brain activity associated to the observation of a robot performing a task has been developed. The design of the protocol does not make any assumptions about the type of ERP response (i.e. the underlying nature of the components of the response). Based on the analysis of the signals, we show that we can automatically distinguish, not only human perceived errors, but also different types of errors (magnitude and laterality). Finally, we show how this detection can be used to learn a different but related task, using classical reinforcement learning.

II. Methodology

A. Protocol Design and Experimentation

The general setting of the experiments was a subject observing a virtual robot on a screen performing a reaching task while the EEG was recorded (Figure 1 (a)). The robot had two degrees of freedom: a revolute joint located at the base of the arm that rotated the full arm and a prismatic joint that made the arm longer. Five different actions moved the robot gripper to each of the five predefined areas (marked as baskets). The subject was instructed to judge the robot motion as follows: (a) a motion towards the central basket is interpreted as a correct operation, (b) a motion towards the baskets placed just on the side (left or right) of the central one is a small operation error, and (c) a motion towards the outside baskets is a large operation error. Figures 1 (b) and (c) show two snapshots of the experiment. Notice that this protocol includes error vs. no errors, plus different levels of operation error and different error locations. The use of a simulated environment allows us to isolate problems (such as robot synchronization, time delays, etc), ensures repeatability among subjects, speeds up the experimentation phase and facilitates the evaluation and characterization of the ERP activity. Two participants participated in the experiments.

The recording session consists of several sequences of actions observed by the subject. Each sequence starts with a five seconds countdown preparing the subject for the operation. A sequence is composed of 10 movements. Each movement starts with the robot at the initial position for one second (Figure 1 (b)), and then switches the arm to one of the five final positions (Figure 1 (c)). After another second, it returns to the initial position and repeats the process. The instantaneous motion between the initial and final positions eliminates the effect of continuous operation and provides a clear trigger on the ERP. A trial consists of five sequences

(a)

(b) (c)

Fig. 1. (a) Experiment setup during the EEG recordings. (b) The initial position of the robot. The task goal is to move the robot gripper to the central basket. (c) Example of an incorrect operation.

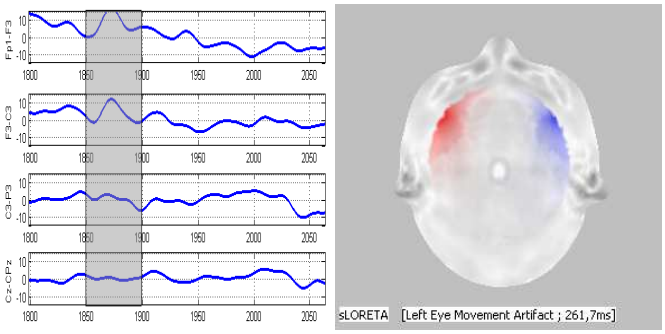
with the five seconds countdown between sequences. The experiment was carried out 12 times (with a relax time of at least 2 minutes after each trial). This process leads on average to 120 ERP responses of each basket, which is the typical amount of samples used in ERP literature to have a good signal to noise ratio using grand averages techniques to study the responses [15].

The instrumentation used to record the EEG was a gTec system (an EEG cap, 32 electrodes, and a gUSBamp amplifier) connected via USB to the computer. The location of the electrodes was selected following previous ERP studies [16] at FP1, FP2, F7, F8, F3, F4, T7, T8, C3, C4, P7, P8, P3, P4, O1, O2, AF3, AF4, FC5, FC6, FC1, FC2, CP5, CP6, CP1, CP2, Fz, FCz, Cz, CPz, Pz and Oz (according to the international 10/20 system). The ground electrode was positioned on the forehead (position FPz) and the reference electrode was placed on the left earlobe.

The EEG was amplified, digitized with a sampling frequency of 256 Hz, and power-line notch-filtered and bandpass-filtered between 0.5 and 10 Hz. As usually done in this type of recordings, a Common Average Reference (CAR) Filter was applied to remove any offset component detected on the signal. The signal recording and processing, the visual application, and the synchronization between the visual stimuli and the EEG were developed under BCI2000 platform [17].

B. Analysis of Artifacts

Artifacts come in many different forms and may have diverse causes. In general, they are non-cerebral potentials (e.g. vigorous motion or eye blinking) that are amplified and



(a) (b)

Fig. 2. (a) Left eye movement artifact (shadowed) recorded in the bipolar channels (from up to down) Fp1-F3, F3-C3, C3-P3, and Cz-CPz. (b) Scalp topoplot in the moment the artifact was maximum (∼260ms after the action performed), showing positive and negative activity in left and right eyes respectively.

may render the EEG uninterpretable. In the present work, it is important to address this issue to: (i) extract conclusions of the underlying mechanisms of human error processing under the present protocol, and (ii) to assure that machine learning algorithms are trained with brain activity samples and not with muscles or other sources of artifacts.

The most common artifacts in these experiments are the eye blink artifacts, muscle action, and chewing and tongue movements. To avoid them, the user was comfortably sat in a chair and instructed not to chew, move the tongue or blink within a sequence of motions of the robot (there was a time between sequences for the user to relax). However, all the experimentation had an artifact that was difficult to avoid given the experimental settings: the lateral eye movement artifact [18]. This artifact is generated by the lateral motion of the eyes, which appeared in many recordings since the subjects tracked the robot motion with the eyes. In general, this artifact is recognizable in the fronto-temporal deviations as sharply contoured potentials that are out phase measured in the frontal electrodes mainly Fp1, Fp2, F3, F4, F7, F8 and temporal T7 and T8. EEG recordings of the central, parietal and occipital lobes are free of this artifact and we confirmed this by visual inspection of the raw EEG.

To address this artifact, all the posterior analyses use only the electrodes over the medial and posterior regions (except channels T7 and T8). Furthermore, we used bipolar recordings as suggested in the ERP literature in order to minimize the effect of the artifacts [19], [15]. Figure 2 shows the EEG in bipolar montage from the anterior to the posterior regions of the brain in one of the participants. Notice how the bipolar montage C3-P3 is already free of this artifact. Furthermore, in the current protocol, the artifact appears on average between 0.2 and 0.3 seconds after the stimulus presentation (time required to saccade to the new robot position) and lasts a maximum of 0.4 seconds.

C. Neurophysiological Response

The objective of this section is to show that the neurophysiological response is coherent for both participants and to characterize the response in terms of the experimental setup. Notice that the objective here is not to characterize the ERP as this would require a much larger number of participants and a complete different type of analysis, as it is usually done in neuropsychology.

To study the responses, we build the averaged ERP waveforms, which is the averaged sum of the individual responses for each condition at each sensor (to improve the signal-to-noise ratio and, as a consequence, filter background noise and occasional artifacts). The averaged ERP waveforms consist of a sequence of components, which are traditionally used to indicate positive-going or negative-going peaks. The sequence of the ERP peaks reflects the flow of information through the brain [19]. Different subjects or conditions usually modify the time and shape of the peaks. In our case, the resulting distribution of the components on the averaged ERP waveforms in Cz (the vertex) of the error vs non-error are coherent in both subjects (Figure 3). This also holds for the ERP waveforms of the left and right error vs non-error and large and small error vs non error.

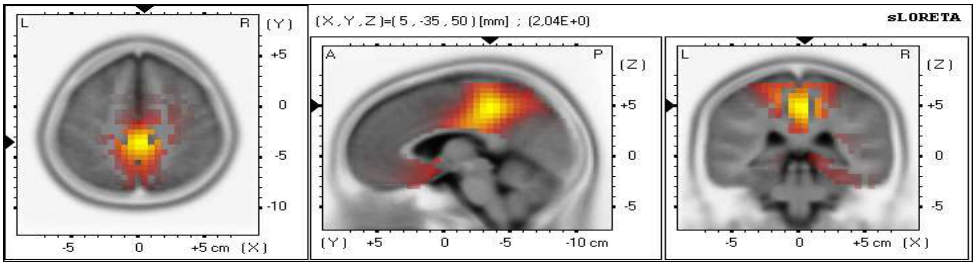
This coherence also appears in the localization of the main brain cortex areas involved in the neural response. For the analysis, we used sLORETA [20], which is an EEG Source Localization technique that estimates the neural generators given the EEG at the surface of the scalp. Note that to solve the inverse problem one needs the 32 electrodes. Since some of them are affected by the artifact, they will affect the solution for earlier components. Thus, since the first 400 ms of the signal are noised with artifacts, the study was performed at the time of the third negativity in Cz (∼510ms and ∼460ms for the first and second participant respectively). With both participants, the main areas active at the negativity studied were Brodmann[1] Areas 5 and 7, which indicates that the same areas are involved in generation of the waveform in both participants (Figure 3).

To characterize the response in terms of the robot actions, a statistical analysis was performed for all the ERPs for each condition. We performed an ANalysis Of Variance (ANOVA) test, since it has been widely used when analyzing differences in ERPs [19], [15], with a significance level of 95% ($p < 0.05$). In the three cases: (i) error versus correct responses, (ii) left errors versus right errors, and (iii) large errors versus small errors, signals are significantly different for the different electrodes at several points in time. For the sake of simplicity, we only display in Figure 3 the ANOVA results in Cz, however other areas have larger statistical differences in other time instants.

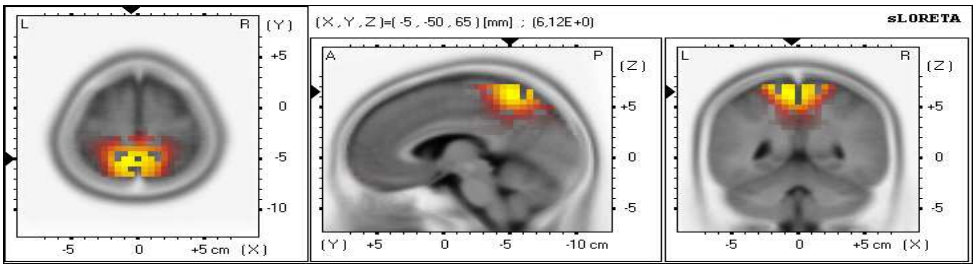
The combination of the previous results allows us to hypothesize about the involvement of the human error monitoring process in our results. Firstly, there is a statistical difference between the human response to the robot correct and incorrect operations. Secondly, the shape of the response in Cz elicited in the incorrect operations is similar to the response of other protocols that involve the human monitoring of errors (see [9] for some examples): all of them have a

[1]The brain cortex can be divided in areas or regions defined according to its cytoarchitecture (the neurons' organization in the cortex). These zones are called Brodmann Areas (BA) [21], and are numerated from 1 to 52.

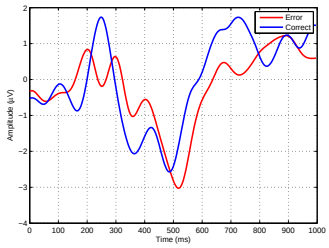
P1 Source Localization



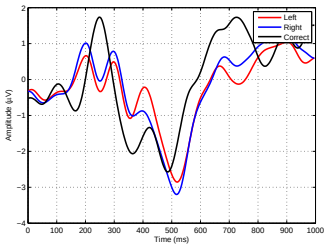
P2 Source Localization



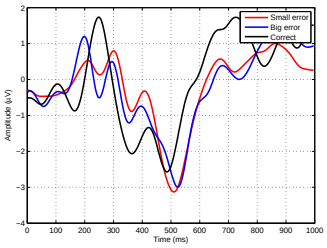
P1 Average Waveform error VS correct



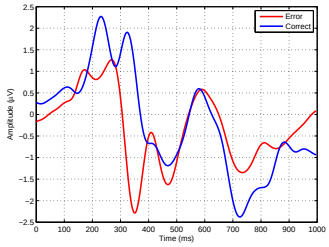
P1 Average Waveform left VS right



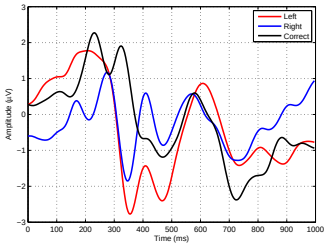
P1 Average Waveform small VS large



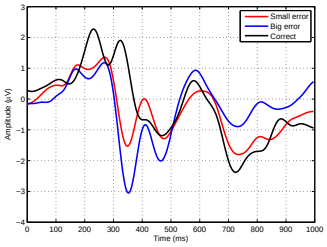
P2 Average Waveform error VS correct



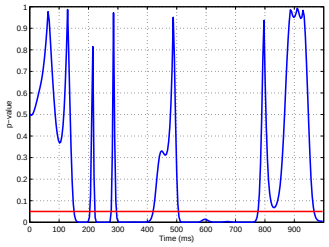
P2 Average Waveform left VS right



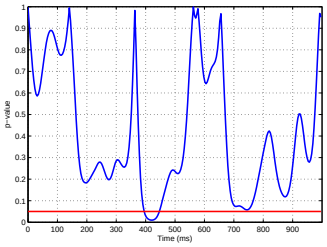
P2 Average Waveform small VS large



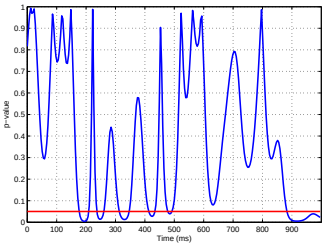
P1 ANOVA error VS correct



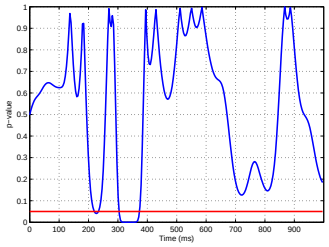
P1 ANOVA left VS right



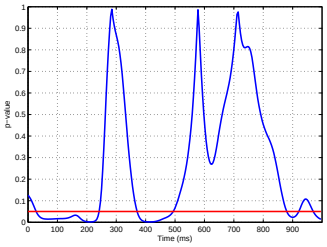
P1 ANOVA small VS large



P2 ANOVA error VS correct



P2 ANOVA left VS right



P2 ANOVA small VS large

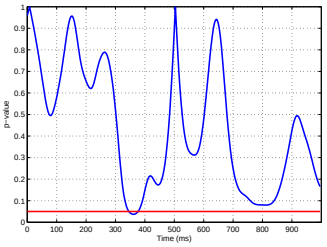


Fig. 3. Source localization, average waveforms and ANOVA analysis of each participant (P1 and P2) in channel Cz. For the ANOVA figure, the vertical axis corresponds to the p-values at each instant of time for the same time window as before. The horizontal red line shows the p-value of 0.05.

sharp positive potential at around 0.3 seconds, followed by a prominent negativity around 0.4 second. Thirdly, in the negativity studied, the activated Brodmann areas are 5 and 7, which are also activated in the interaction error [9] in the late components. Their hypothesis is that these associative areas (somatosensory association cortex) could be related to the fact that the subject becomes aware of the error. This also agrees with findings using other types of errors such as reaction errors [22][2]. All these results push forward the hypothesis that we detect an ERP that is related with the human error monitoring of the robot operation.

D. Pattern recognition

The analysis of the EEG signals of the previous section revealed that there are statistically significant differences in the brain activity when observing correct or wrong actions and for the different conditions of error. In order to provide the reward signal to a RL algorithm, it is necessary to classify single trials of these error conditions online.

The pattern recognition is a supervised learning module that is trained to recognize the ERP responses. It requires two steps. The first one is the feature extraction. Firstly, the ANOVA analysis was performed over all the bipolar channels in the medial and posterior regions. The channels with more statistical difference were selected: C3-P3, C4-P4, P3-O1, P4-O2, CP5-P3, CP6-P4, CP5-P7, CP6-P8, CP1-Pz, CP2-Pz, Fz-FCz, FCz-Cz, Cz-CPz, CPz-Pz, and Pz-Oz. We used as features the RAW signal within the time window 0.15-0.7 seconds, filtered with a CAR filter and a bandpass filter of 0.5-10Hz (as mentioned previously), and then subsampled to 64Hz. Thus, the feature vector was the concatenation of all the selected channels within the window previously defined, giving a feature vector length of 540.

The second step is the classification algorithm. We chose AdaBoost classification algorithm [23]. This classifier has the advantage of being a meta-classifier, i.e, it makes use of several weak classifiers and assigns weights to them. This technique has been successfully used in several applications [24]. As weak classifier, we chose the Functional Decision Tree [25], which allows to use linear combinations of attributes, due to the multi-variate nature of the EEG data. We experimentally verified that this combination achieves good classification performances.

E. Reinforcement Learning

The final step of this work is to show how error related potentials detected online from EEG can be used for a robot learning task within a reinforcement learning context. The main idea of RL is that an agent (a robot in our case) learns by interacting with the environment from a signal r that rewards or penalizes its behavior. The standard framework for RL problems is a Markov Decision Process defined by the tuple $\{S, A, P, r, \gamma\}$ where S represents the state-space, A represents the action-space and $P : S \times A \rightarrow S$ are the transition probabilities from state s to state s' when executing a particular action a. The function $r : S \times A \rightarrow \mathcal{R}$ defines the reward obtained by the agent when executing an action at a particular state. Finally, $\gamma \in [0..1]$ is a discount factor.

The goal of RL is to obtain a policy $\pi : S \rightarrow A$ that maps each state to the action that maximizes the accumulated reward $R_k = \sum_{k=0}^{\infty} \gamma^t r_{k+1}$. This is done by maximizing the expected reward, called value function, conditioned in the state s and policy π,

$$V^{\pi}(s) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{k+1} \mid \pi, s_0 = s\right\}.$$

The optimal function V^* satisfies the Bellman equation,

$$V^*(s) = max_{a \in A}\left[r(s,a) + \sum_{s' \in S} P(s,a,s')V^*(s')\right].$$

It has been proved that at least an optimal policy π^* exists for any finite MDP. There are several algorithms to compute the optimal policy. For the type of discrete tasks described in Section II-A, we used the standard Q-Learning algorithm [2] which uses the Q-function

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in S} P(s,a,s')max_{a' \in A}Q^*(s',a').$$

Q-learning estimates the optimal Q^* function from empirical data. It does not require to know the transition probabilities P. At each point in time, the agent is at a particular state s_k, executes an action a_k that results in a new state s_{k+1} and obtains reward $r_{k+1}(s_k, a_k)$. Based on this observed transition and reward the Q-function is updated using

$$Q_{k+1}(s_k,a_k) = Q_k(s_k,a_k) + \alpha_k\left[r_{k+1}(s_k,a_k) + \gamma \max_{a' \in A} Q_k(s_{k+1},a') - Q_k(s_k,a_k)\right],$$

where $Q_k(\cdot,\cdot)$ is the current estimate of the Q-function, and α is the learning rate.

During learning, it is necessary to choose the next action to execute. We will use an ε-greedy policy. This type of policies choose the best action (obtained from the current policy) $(100-\varepsilon)\%$ of the times and will select an exploration strategy by choosing a random action $\varepsilon\%$ of the times, normally a low value (around 10%) to take more into account the acquired knowledge.

It is worth noting that in real settings one cannot train a classifier for each task and compute the reward from its output, since this would require to label the data and will make the EEG signal redundant. In practice, the classifier has to be trained on a set of related examples and this knowledge has to be transferred to the new task. Although transfer learning is an emergent research area [26], in our experiments we will simply vary the task goal (i.e the reward values) and keep the same Markov Process to study the invariance of brain activity in this context.

[2]We cannot provide results for the activation of the Anterior Cingulate Cortex (ACC) which is involved in the error processing [9]. Unfortunately, this area is relevant in the early components of the response and, therefore, the EEG signals are contaminated by artifacts.

III. RESULTS

In this section we present the results for the single trial automatic classification of errors following Section II-D. The results focus directly on the classification between the correct and all the different incorrect operations, including magnitude and laterality of the error (5 classes). This is because this information is very valuable for a reinforcement learning task, since the rewards could be set according to the error. In a second step, we show how it is possible to use the previous classifier to learn, using the RL method described in Section II-E, the correct action arbitrary selected by the user. The main difficulty here is that this task is related but different from the previous one, and thus the classifier has to transfer prior knowledge (i.e generalize) from the training examples to the new task.

A. Pattern Recognition performance

In this section, we analyze the results of single trial automatic classification of the EEG signal. In order to determine the accuracy of the classifier and make use of all the data, we used a ten-fold cross-validation strategy. The classification was made using five different classes corresponding to the grasping areas (baskets) of the experiment. These classes are labeled as (from left to right baskets): Left-2 (large left error), Left-1 (small left error), Correct (correct responses), Right-1 (small right error), and Right-2 (large right error). The results for each participant (P1 and P2) are shown in tables I and II. Each column represents the real class, whereas each row show the actual classification percentages for each class, thus having the correct classification always on the diagonal.

TABLE I
PATTERN RECOGNITION PERFORMANCE, P1

Left-	Left-2	Left-1	Correct	Right-1	Right-2
Left-2	74.17%	20.83%	3.33%	0.83%	0.83%
Left-1	20.00%	68.33%	9.17%	2.50%	0.00%
Correct	2.50%	7.50%	79.17%	8.33%	2.50%
Right-1	0.83%	6.67%	5.00%	68.33%	19.17%
Right-2	1.67%	0.83%	5.83%	21.67%	70.00%

TABLE II
PATTERN RECOGNITION PERFORMANCE, P2

	Left-2	Left-1	Correct	Right-1	Right-2
Left-2	62.50%	25.83%	2.50%	2.50%	6.67%
Left-1	23.33%	61.67%	10.00%	4.17%	0.83%
Correct	2.50%	5.00%	78.33%	5.83%	8.33%
Right-1	1.67%	5.83%	7.50%	60.00%	25.00%
Right-2	5.83%	8.33%	4.17%	23.33%	58.33%

The results vary slightly depending on the subject. The recognition rate for participant 1 is always around 70% and almost reaches 80% for the correct case. Results for participant 2 achieve the same result for correct actions, but are closer to 60% when recognizing different levels of error. An interesting result is that the misclassifications tend to be grouped around similar errors for both participants. For instance, the largest error percentages, around 20% for participant 1 and 25% for participant 2, occur between the detection of Left-2 and Left-1 and the detection of Right-2 and Right-1. In other words, detecting left and right errors achieved an accuracy over 90%. Regarding the differences between small and large errors, the classifier performance degrades a little. The percentages vary again for each subject. Results for participant 1 were 72.08% for large errors and 68.33% for small errors while for participant 2 they were 60.42% and 60.83%, respectively. The results of the AdaBoost classifier suggest that it is possible to differentiate between correct and wrong actions. The misclassification rate, being still non negligible, is low enough to produce reward signals that can be exploited by a reinforcement learning algorithm. Interestingly, despite we are in a discrete setup, there are also strong indications that it is possible to recover additional information related to the type of error. It seems that magnitude and directionality information are present in the signal and could be potentially exploited for the learning task in continuous-state spaces. However, other classifiers may be needed when distinguishing between small and large errors due to the degradation in performance for the selected classifier.

B. RL application

In order to analyze the practical potential of this approach, we have applied it to a simple Reinforcement Learning task. This new task is based on the experimental setup described in Section II-A. However, in this case, the participant is instructed to freely select one grasping area to set the correct operation of the robot. The objective was to allow the robot to learn the correct action (motion towards the selected area or basket) using a RL algorithm.

The first participant selected the basket 1, whereas the second participant selected the basket 4. We repeated the experimental protocol for each subject as described in subsection II-A obtaining 120 ERPs of each basket. Since each subject selected the basket freely, we needed to relabel the classes according to their selection. Labels for participant 1 were (from left to right baskets): Left-3, Left-2, Left-1, Correct, and Right-1. Notice that in this case we have larger errors (Left-3) than in the first experiment. Labels for participant 2 were (from left to right baskets): Correct, Right-1, Right-2, Right-3, and Right-4. Again, we have larger errors (Right-3 and Right-4) than in the first experiment.

Despite we changed the experiment, we will still use the classifier from the experiment with marked baskets to classify the signals obtained in the new experiment. Tables III and IV show the results for participant 1 (P1) and participant 2 (P2). Since we have different experiments, the labels do not perfectly match. The rows contain the actual errors according to each participant choice of correct basket, whereas the columns still show the same classes of Section III-A. The key issue here is that, to apply an RL algorithm and learn a new task, we need to classify signals based on a classifier trained on a different task, but that can transfer some invariant information, in our particular case, the correct basket and

the spatial relations to the others. In other words, we expect that a Left-1 error will still be a Left-1 error. In addition to this, new classes appear (Left-3, Right-3 and Right-4), and we also expect that the classifier will select classes that keep some relations. For instance, Left-3 errors should be matched to Left-2 errors (due to directionality).

TABLE III
PATTERN RECOGNITION PERFORMANCE EXPERIMENT 2, P1

	Left-2	Left-1	Correct	Right-1	Right-2
Left-3	62.50%	27.50%	5.00%	4.17%	0.83%
Left-2	28.33%	55.83%	9.17%	5.00%	1.67%
Left-1	4.17%	45.83%	35.00%	13.33%	1.67%
Correct	0.00%	3.33%	52.50%	31.67%	12.50%
Right-1	0.00%	0.00%	8.33%	31.67%	60.00%

TABLE IV
PATTERN RECOGNITION PERFORMANCE EXPERIMENT 2, P2

	Left-2	Left-1	Correct	Right-1	Right-2
Correct	14.17%	25.00%	48.33%	4.17%	8.33%
Right-1	12.50%	15.83%	45.83%	20.00%	5.83%
Right-2	7.50%	3.33%	23.33%	40.83%	25.00%
Right-3	6.67%	1.67%	8.33%	41.67%	41.67%
Right-4	12.50%	4.17%	15.00%	24.17%	44.17%

The classification results of Tables III and IV show that the correct action was detected with a performance of 52.5% and 48.33% respectively for each participant. Notice that the diagonal does no longer contain correct associations, but one has to look for the same label for rows and columns. As in the previous case, the performance of the classifier was better for the first participant. This effect was also amplified by the fact that participant 2 chose the leftmost basket. Further analysis of the data show that confusion among classes still keeps some coherent structure and usually wrong classifications occur among similar classes. Considering small errors (Left-1 and Right-1) of participant 1, we obtain accuracies of 45.83% and 31.67% respectively. For the Right-1 case, it was detected more frequently (60%) as a Right-2 error. Another example for this participant is Left-3 errors, which are frequently detected as Left-2 errors (62.5%), which is a good result for a previously unknown class. These results also extend for participant 2, detecting Right-3 and Right-4 errors usually as Right-2 errors (41.67% and 44.17% respectively).

Finally, we have used the Q-Learning algorithm to determine the correct action (basket) selected by the user from the EEG recorded activity. The system started with the same Q-functions values. Actions were selected according to the ε-greedy policy described in Section II-E. The rewards associated to each executed action were computed based on the class assigned by the classifier: -1 for large errors, -0.5 for small errors, and $+1$ for correct actions.

Due to the low detection ratios, Q-learning did not always converge to the correct basket. We executed 20 times the Q-learning algorithm with the previously classified data.

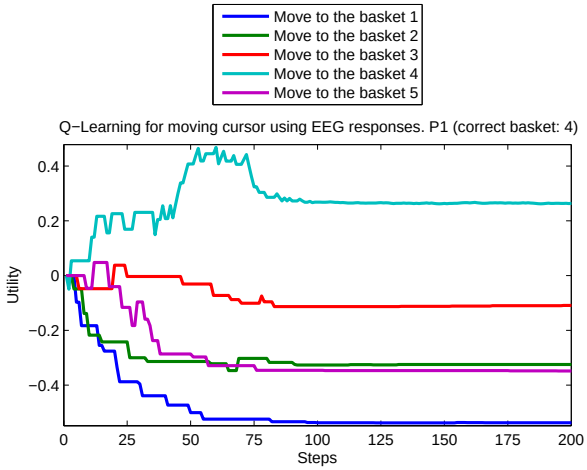


Fig. 4. Q-Learning results of executing each action for participant 1 using bipolar channels.

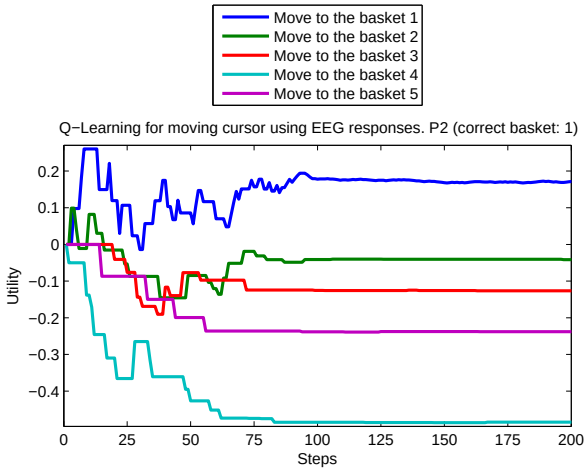


Fig. 5. Q-Learning results of executing each action for participant 2 using bipolar channels.

For participant 1, 92% of the executions discovered the correct basket. Convergence usually took around 70 steps. For participant 2, Q-learning converged in 75% of the cases and required around 100 steps on average. Figures 4 and 5 show examples where it converged for each participant.

In summary, we have shown that it is possible to apply RL using EEG based reward signals. Although the current setup is very simple, it illustrates some of the main issues to be considered in this type of applications. In particular, we would like to stress that the results show an implicit transfer of knowledge between two different tasks. The ability of the classifier to (still poorly) generalize between signals corresponding to different tasks is a very encouraging and promising result and an indication of the feasibility of EEG-based reinforcement learning.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have studied the use of brain activity to create reward signals for reinforcement learning. We

have introduced a new protocol to study the ERP activity associated to the evaluation of a task. The results show that there exist statistically significant differences in the grand averages of the signal, not only between error and correct actions, but also among different degrees of errors. Using boosting techniques, we have been able to detect single trials of different types of errors automatically. Finally, the system was able to learn the correct action (i.e. basket) selected freely by the user using the classifier trained on a different one.

There are plenty of opportunities for future work. First, we need to better characterize the components of the brain activity associated to the proposed protocol. Despite this paper has presented a proof-of-concept experiment with two participants, a component characterization will require to conduct further experiments with a larger number of participants to verify the hypotheses about the ERP nature of the recorded brain activity. Second, the results suggest that it is possible to obtain information about the correct execution of the task that goes beyond simple error vs. non error classification. This information would be extremely useful to perform reinforcement learning in more realistic and complex robot scenarios.

REFERENCES

[1] S. Schaal, A. Ijspeert, and A. Billard, "Computational approaches to motor learning by imitation," *Phil. Trans. of the Royal Society of London: Series B, Biological Sciences*, vol. 358, no. 1431, 2003.

[2] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[3] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural Networks*, vol. 21, no. 4, pp. 682–697, 2008.

[4] C. Guger, W. Harkam, C. Hertnaes, and G. Pfurtscheller, "Prosthetic control by an EEG-based braincomputer interface (BCI)," *Proceedings of AAATE 5th European conference for the advancement of assistive technology*, 1999.

[5] J.d.R. Millán, F. Renkens, J. Mouriño, and W. Gerstner, "Noninvasive Brain-Actuated Control of a Mobile Robot by Human EEG," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, June 2004.

[6] I. Iturrate, J. Antelis, A. Kuebler, and J. Minguez, "Non-Invasive Brain-Actuated Wheelchair based on a P300 Neurophysiological Protocol and Automated Navigation," *IEEE Transactions on Robotics*, vol. 25, no. 3, pp. 614–627, 2009.

[7] C. Escolano, J. Antelis, and J. Minguez, "Human Brain-Teleoperated Robot between Remote Places," *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[8] H.G Vaughan, *The relationship of brain activity to scalp recordings of event-related potentials*, pp. 45–75, Washington, D.C.: U.S. Government Printing Office., 1969.

[9] P.W. Ferrez and J.d.R. Millan, "Error-related eeg potentials generated during simulated brain-computer interaction," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 923–929, March 2008.

[10] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, "ERP components on reaction errors and their functional significance: A tutorial," *Biological Psychology*, vol. 51, pp. 87–107, 2000.

[11] S. Nieuwenhuis, C.B. Holroyd, N. Mola, and M.G.H. Coles, "Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance," *Neuroscience and Biobehavioral Reviews*, vol. 28, pp. 441448, 2004.

[12] H.T. van Schie, R.B. Mars, M.G.H Coles, and H. Bekkering, "Modulation of activity in medial frontal and motor cortices during error observation," *Neural Networks*, vol. 7, pp. 549–554, 2004.

[13] B. Dal Seno, *Toward An Integrated P300- And ErrP-Based Brain-Computer Interface*, Ph.D. thesis, Politecnico di Milano, 2009.

[14] R. Chavarriaga, P.W. Ferrez, and J.d.R. Millan, "To Err is Human: Learning from Error Potentials in Brain-Computer Interfaces ," *1st International Conference on Cognitive Neurodynamics*, 2007.

[15] T. Handy, Ed., *Event-Related Potentials A Methods Handbook*, The MIT Press, 2005.

[16] P.W. Ferrez, *Error-Related EEG Potentials in Brain-Computer Interfaces*, Ph.D. thesis, École Polytechnique Fédérale de Laussane, 2007.

[17] G. Shalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw, "BCI2000: A General-Purpose Brain-Computer Interface (BCI) System," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, May 2004.

[18] A.J. Rowan and E. Tolunsky, *Primer of EEG: With A Mini-Atlas*, Butterworth-Heinemanns, 2003.

[19] S.J. Luck, *An Introduction to the Event-Related Potential Technique*, The MIT Press, 2005.

[20] R.D. Pascual-Marqui, "Standardized low resolution brain electromagnetic tomography (sLORETA): Technical details.," *Methods and Findings in Experimental and Clinical Pharmacology*, pp. 5–12, 2002.

[21] K. Brodmann, *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*, Johann Ambrosius Barth Verlag, 1909.

[22] S. Nieuwenhuis, K.R. Ridderinkhof, J. Blom, G.P.H. Band, and A. Kok, "Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task," *Psychophysiology*, vol. 38, pp. 752–760, 2001.

[23] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *European Conference on Computational Learning Theory*, pp. 23–37, 1995.

[24] O.M. Mozos, *Semantic Place Labeling with Mobile Robots*, Ph.D. thesis, Dept. of Computer Science, University of Freiburg, July 2008.

[25] J. Gama, "Functional trees," *Machine Learning*, vol. 55, no. 3, pp. 219–250, November 2004.

[26] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *Available online: http://www.cse.ust.hk/ sinnopan/publications/TLsurvey_0822.pdf*, 2009.